Bayesian inference & linguistic parameters

Lisa Pearl
University of California, Irvine

August 5, 2017:
Norwegian Summer Institute on Language & Mind
University of Oslo

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]
Today’s Plan:
Bayesian inference & linguistic parameters

I. Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

II. Parameters & overhypotheses

III. Structure dependence
Today’s Plan:
Bayesian inference & linguistic parameters

I. Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]
A Bayesian model assumes the learner has some space of hypotheses $H$...
A Bayesian model assumes the learner has some space of hypotheses $H$, each of which represents a possible explanation for how the data $D$ in the data intake were generated.
A Bayesian model assumes the learner has **some space of hypotheses** $H$, each of which represents a possible explanation for how **the data** $D$ in the data intake were generated.

**Example parameter:**
Subject drop

**English:** -subj-drop
Requires Subject to be overt

- Subject Verb
  - They drink

- Verb Drink

“**They drink**”
Bayesian reasoning

A Bayesian model assumes the learner has some space of hypotheses $H$, each of which represents a possible explanation for how the data $D$ in the data intake were generated.

Example parameter:
Subject drop

![Diagram showing two hypotheses, h1 and h2, with h2 containing h1, which in turn contains the parameter Subject Verb: -subj-drop.](image)
Bayesian reasoning

A Bayesian model assumes the learner has some space of hypotheses $H$, each of which represents a possible explanation for how the data $D$ in the data intake were generated.

Example parameter:
Subject drop  

Spanish: +subj-drop  
Allows Subject to be overt or dropped

✓ Subject Verb
Ellos  beben
they  drink-3rd-pl

✓ Verb
Beben
drink-3rd-pl

“This drink”
Bayesian reasoning

Given $D$, the modeled child’s goal is to determine the probability of each possible hypothesis $h \in H$, written as $P(h|D)$ - the *posterior* for that hypothesis.
Bayesian reasoning

This depends on a few different aspects (which have their own probabilities).

\[ P(h|D) = \frac{P(D|h)*P(h)}{P(D)} \]
Bayesian reasoning

$P(D|h)$ represents the *likelihood* of the data $D$ given hypothesis $h$, and describes how compatible that hypothesis is with the data.

$$P(h|D) = \frac{P(D|h)*P(h)}{P(D)}$$
Bayesian reasoning

$P(D|h)$ represents the likelihood of the data $D$ given hypothesis $h$, and describes how compatible that hypothesis is with the data.

$$P(h|D) = \frac{P(D|h) * P(h)}{P(D)}$$

What if the data intake contained both data point types?
Bayesian reasoning

$P(D | h)$ represents the *likelihood* of the data $D$ given hypothesis $h$, and describes how compatible that hypothesis is with the data.

$$P(h | D) = \frac{P(D | h) \cdot P(h)}{P(D)}$$

$P(D | h_1) = 1^*$

-subj-drop can account for Subject Verb.

Because this is the only data point it can generate in this scenario, the probability of generating $D$ is $1/1 = 1$. 
Bayesian reasoning

P \( (D|h) \) represents the *likelihood* of the data D given hypothesis h, and describes how compatible that hypothesis is with the data.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

P(\(D \mid h1\)) = 1 \* 0 = 0

-subj-drop can’t account for Verb alone.
Bayesian reasoning

P (D|h) represents the *likelihood* of the data D given hypothesis h, and describes how compatible that hypothesis is with the data.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

\[
P(D | h1) = 0
\]

\[
P(D | h2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}
\]

+subj-drop can account for both data points. The probability of generating each one is 1/2.
Bayesian reasoning

P(D|h) represents the *likelihood* of the data D given hypothesis h, and describes how compatible that hypothesis is with the data.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

What if the data intake contained only this data point type?
Bayesian reasoning

P(D|h) represents the *likelihood* of the data D given hypothesis h, and describes how compatible that hypothesis is with the data.

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

P(D | h1) = 1

-subj-drop can account for it

Because this is the only data point it can generate in this scenario, the probability of generating D is 1/1 = 1.
Bayesian reasoning

P(D|h) represents the *likelihood* of the data D given hypothesis h, and describes how compatible that hypothesis is with the data.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

\[P(D \mid h_1) = 1\]
\[P(D \mid h_2) = \frac{1}{2}\]

+subj-drop can generate it too.

Because +subj-drop can generate two data points, the probability of generating this data point is 1/2.
Bayesian reasoning

P (h) represents the *prior* of the hypothesis h, and represents the probability of the hypothesis before any data have been encountered. Intuitively, this corresponds to how plausible the hypothesis is, irrespective of any data.

$$P(h|D) = \frac{P(D|h)*P(h)}{P(D)}$$

This is often where considerations about the complexity of the hypothesis will be implemented.
Bayesian reasoning

P (h) represents the *prior* of the hypothesis h, and represents the probability of the hypothesis before any data have been encountered. Intuitively, this corresponds to how plausible the hypothesis is, irrespective of any data.

\[
P(h|D) = \frac{P(D|h)*P(h)}{P(D)}\]

If there’s no reason to consider one hypothesis more complex than another, the hypotheses will typically receive **uniform** probability (all of them have the same probability).

This is typically 1 over the total hypotheses available.
Bayesian reasoning

$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$

uniform probability

$P(h_1) = 1/2$
$P(h_2) = 1/2$

$P(h)$ represents the *prior* of the hypothesis $h$, and represents the probability of the hypothesis before any data have been encountered. Intuitively, this corresponds to how plausible the hypothesis is, irrespective of any data.
Bayesian reasoning

P(D) represents the probability of the data irrespective of any hypothesis. It serves as a normalizing factor so that the posterior probabilities sum to 1.

\[
P(h|D) = \frac{P(D|h)*P(h)}{P(D)}
\]
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} = \frac{\sum_{h' \in H} P(D|h) \cdot P(h)}{\sum_{h' \in H} P(D|h') \cdot P(h')} \]
Bayesian reasoning

\( P(D) \) is calculated by summing over all possible hypotheses the following:

the likelihood of the hypothesis \( \times \) the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} = \frac{\sum_{h' \in H} P(D|h') \times P(h')}{P(D)}
\]
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} = \frac{P(D|h) \times P(h)}{\sum_{h' \in H} P(D|h') \times P(h')} \]

\text{P(D)} \text{ is calculated by summing over all possible hypotheses the following:}

the \textit{likelihood} of the hypothesis * the \textit{prior} of the hypotheses.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

\[ = \frac{P(D|h) \cdot P(h)}{\sum_{h' \in H} P(D|h') \cdot P(h')} \]

\[ \propto P(D|h) \cdot P(h) \]

Because we often only care about how one hypothesis compares to another, calculating \( P(D) \) can be skipped over.

Why is this so?
Bayesian reasoning

\( P(D) \) is calculated by summing over all possible hypotheses the following:
the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} = \frac{P(D|h) \times P(h)}{\sum_{h' \in H} P(D|h') \times P(h')}
\]

\( \propto P(D|h) \times P(h) \)
Bayesian reasoning

\[ P(D) \text{ is calculated by summing over all possible hypotheses the following:} \]
\[ \text{the likelihood of the hypothesis } \ast \text{ the prior of the hypotheses.} \]

\[
P(h|D) = \frac{P(D|h) \ast P(h)}{P(D)}
\]

\[
= \frac{P(D|h) \ast P(h)}{\sum_{h' \in H} P(D|h') \ast P(h')}
\]

\[
\propto P(D|h) \ast P(h)
\]

Likelihoods

\[ P(D | h1) = 1 \]
\[ P(D | h2) = 1/2 \]
Bayesian reasoning

\( P(D) \) is calculated by summing over all possible hypotheses the following:

the likelihood of the hypothesis \( \times \) the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} = \frac{P(D|h) \times P(h)}{\sum_{h' \in H} P(D|h') \times P(h')}
\]

Data \( D \)

Subject Verb

likelihoods

\[
P(D \mid h1) = 1
\]

\[
P(D \mid h2) = \frac{1}{2}
\]

priors

\[
P(h1) = \frac{1}{2}
\]

\[
P(h2) = \frac{1}{2}
\]
Bayesian reasoning

$P(D)$ is calculated by summing over all possible hypotheses the following:

the **likelihood** of the hypothesis * the **prior** of the hypotheses.

$$P(h|D) = \frac{P(D|h) \times P(h)}{P(D)}$$

$$\propto P(D|h) \times P(h)$$

---

Data $D$

Subject Verb

**likelihoods**

$P(D | h_1) = 1$

$P(D | h_2) = 1/2$

**priors**

$P(h_1) = 1/2$

$P(h_2) = 1/2$

**likelihood * prior**

$P(D | h_1) \times P(h_1) = 1 \times 1/2 = 1/2$

$P(D | h_2) \times P(h_2) = 1/2 \times 1/2 = 1/4$
Bayesian reasoning

\[ P(D) \text{ is calculated by summing over all possible hypotheses the following:} \]

the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \times P(h)}{P(D)}
\]

\[
= \frac{P(D|h) \times P(h)}{\sum_{h' \in \mathcal{H}} P(D|h') \times P(h')}
\]

Data \( D \)

Subject Verb

likelihood * prior

P(D | h1) * P(h1) = 1/2

P(D | h2) * P(h2) = 1/4

sum 3/4
Bayesian reasoning

\[ P(D) \] is calculated by summing over all possible hypotheses the following:

the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

\[ = \frac{P(D|h) \cdot P(h)}{\sum_{h' \in H} P(D|h') \cdot P(h')}\]

\[ \propto P(D|h) \cdot P(h) \]

Data \( D \)

Subject Verb

likelihood * prior

\[ P(D \mid h1) \cdot P(h1) = \frac{1}{2} \]

\[ P(D \mid h2) \cdot P(h2) = \frac{1}{4} \]

sum \( \frac{3}{4} \)

\[ P(h1 \mid D) = \]
Bayesian reasoning

P(D) is calculated by summing over all possible hypotheses the following:

the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \times P(h)}{P(D)}
\]

= \frac{P(D|h) \times P(h)}{\sum_{h' \in H} P(D|h') \times P(h')}

\[
\propto P(D|h) \times P(h)
\]

Data D

Subject Verb

likelihood * prior

P(D | h1) * P(h1) =

P(D | h2) * P(h2) = 1/4

sum

P(h1 | D) = \frac{1/2}{3/4} = 2/3
Bayesian reasoning

$P(D)$ is calculated by summing over all possible hypotheses the following:

the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} = \frac{P(D|h) \cdot P(h)}{\sum_{h' \in H} P(D|h') \cdot P(h')}
\]

\[
\propto P(D|h) \cdot P(h)
\]

Data $D$

Subject Verb

likelihood * prior

$P(D \mid h1) \cdot P(h1) = \frac{1}{2}$

$P(D \mid h2) \cdot P(h2) = \frac{1}{3}$

sum

$P(h1 \mid D) = \frac{2}{3}$ \quad $P(h2 \mid D) = \frac{1/4}{3/4} = \frac{1}{3}$
Bayesian reasoning

$P(D)$ is calculated by summing over all possible hypotheses the following:

the likelihood of the hypothesis $\times$ the prior of the hypotheses.

$P(h|D) = \frac{P(D|h) \times P(h)}{P(D)}$

$$= \frac{P(D|h) \times P(h)}{\sum_{h' \in H} P(D|h') \times P(h')}$$

Data $D$

Subject Verb

likelihood $\times$ prior

$P(D | h1) \times P(h1) = 1/2$

$P(D | h2) \times P(h2) = 1/4$

sum $3/4$

$P(h1 | D) = 2/3 \quad P(h2 | D) = 1/3$

Conclusion: $h1$ is now twice as likely as $h2$
Bayesian reasoning

\[ P(D) \text{ is calculated by summing over all possible hypotheses the following:} \]

the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h)*P(h)}{P(D)}
\]

\[
= \frac{P(D|h)*P(h)}{\sum_{h' \in \mathcal{H}} P(D|h')*P(h')}
\]

Data D
likelihood * prior
P(D | h1) * P(h1) = 1/2
P(D | h2) * P(h2) = 1/4

Conclusion: h1 is now twice as likely as h2
Bayesian reasoning

\[ P(D) \text{ is calculated by summing over all possible hypotheses the following:} \]

the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

\[
= \frac{P(D|h) \cdot P(h)}{\sum_{h' \in H} P(D|h') \cdot P(h')}
\]

Data \( D \) is compatible with both hypotheses.

- **h1** is now twice as likely as **h2** even though the data point seen was ambiguous — it was compatible with both hypotheses.
Bayesian reasoning

\( P(D) \) is calculated by summing over all possible hypotheses the following:
the likelihood of the hypothesis * the prior of the hypotheses.

\[
P(h | D) = \frac{P(D | h) \cdot P(h)}{P(D)} = \frac{P(D | h) \cdot P(h)}{\sum_{h' \in H} P(D | h') \cdot P(h')}
\]

\( P(D | h_1) \cdot P(h_1) = \frac{1}{2} \)
\( P(D | h_2) \cdot P(h_2) = \frac{1}{4} \)

The reason is because of the likelihood. \( h_1 \) is a better fit for the data -- it doesn't predict other data like \( h_2 \) does.
Bayesian reasoning

$$P(h|D) = \frac{P(D|h) * P(h)}{P(D)}$$

We have behavioral evidence that very young children reason in a way that leads to similar conclusions when given this kind of scenario.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Infants were trained on data from an artificial language, which consisted of words following a certain pattern.

\[ P(h|D) = \frac{P(D|h) * P(h)}{P(D)} \]
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artiﬁcial language study

The infant’s job: determine the generalization that describes the pattern for words of the artificial language.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Marcus et al. (1999) found that very young infants will notice that words made up of 3 syllables follow a pattern that can be represented as AAB or ABA.

Data D

h1

h2
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Marcus et al. (1999) found that very young infants will notice that words made up of 3 syllables follow a pattern that can be represented as AAB or ABA.

Example:
A syllables = le, wi
B syllables = di, je
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

**AAB or ABA**

A syllables = le, wi  
B syllables = di, je

AAB language words: 
leledi, leleje, wiwidi, wiwije

ABA language words: 
ledile, lejele, widiwi, wijewi
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

**Gerken 2006, 2010**
artificial language study

**AAB or ABA**

AAB language words:
leledi, leleje, wiwidi, wiwije

ABA language words:
ledile, lejele, wiwiwi, wijewi

What kind of generalization would children make if they were given particular kinds of data from these same artificial languages?

Data D
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) * P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

<table>
<thead>
<tr>
<th>di</th>
<th>je</th>
<th>li</th>
<th>we</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>leledi</td>
<td>leleje</td>
<td>leleli</td>
</tr>
<tr>
<td>wi</td>
<td>wiwidi</td>
<td>wiwijе</td>
<td>wiwili</td>
</tr>
<tr>
<td>ji</td>
<td>jijidi</td>
<td>jijije</td>
<td>jijili</td>
</tr>
<tr>
<td>de</td>
<td>dededi</td>
<td>dedeje</td>
<td>dedeli</td>
</tr>
</tbody>
</table>

Infants only see a subset of the language

Data D
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

<table>
<thead>
<tr>
<th></th>
<th>di</th>
<th>je</th>
<th>li</th>
<th>we</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>leledi</td>
<td>leleje</td>
<td>leleli</td>
<td>lelewe</td>
</tr>
<tr>
<td>wi</td>
<td>wiwidi</td>
<td>wiwije</td>
<td>wiwili</td>
<td>wiwiwe</td>
</tr>
<tr>
<td>ji</td>
<td>jijidi</td>
<td>jijije</td>
<td>jijili</td>
<td>jijiwe</td>
</tr>
<tr>
<td>de</td>
<td>dededi</td>
<td>dedeje</td>
<td>dedeli</td>
<td>dedewe</td>
</tr>
</tbody>
</table>

Experimental condition

Training on four word types:
leledi, wiwidi, jijidi, dededi

Consistent with both a less-general hypothesis (h1) and a more-general hypothesis (h2).
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

<table>
<thead>
<tr>
<th></th>
<th>di</th>
<th>je</th>
<th>li</th>
<th>we</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>leledi</td>
<td>leleje</td>
<td>leleli</td>
<td>lelewe</td>
</tr>
<tr>
<td>wi</td>
<td>wiwidi</td>
<td>wiwije</td>
<td>wiwili</td>
<td>wiwiwe</td>
</tr>
<tr>
<td>ji</td>
<td>jijidi</td>
<td>jijije</td>
<td>jijili</td>
<td>jijiwe</td>
</tr>
<tr>
<td>de</td>
<td>dededi</td>
<td>dedeje</td>
<td>dedeli</td>
<td>dedewe</td>
</tr>
</tbody>
</table>

**Control condition**
Training on four word types:
leledi, wiwije, jijili, dedewe

Consistent only with the more-general hypothesis (h2).
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

<table>
<thead>
<tr>
<th></th>
<th>di</th>
<th>je</th>
<th>li</th>
<th>we</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>leledi</td>
<td>leleje</td>
<td>leleli</td>
<td>lelewe</td>
</tr>
<tr>
<td>wi</td>
<td>wiwidi</td>
<td>wiwije</td>
<td>wiwili</td>
<td>wiwiwe</td>
</tr>
<tr>
<td>ji</td>
<td>jijidi</td>
<td>jijije</td>
<td>jijili</td>
<td>jijiwe</td>
</tr>
<tr>
<td>de</td>
<td>dededi</td>
<td>dedeje</td>
<td>dedeli</td>
<td>dedewe</td>
</tr>
</tbody>
</table>

Control condition

Training on four word types:
leledi, wiwije, jijili, dedewe

Consistent only with the more-general hypothesis (h2).

This control condition is used to see what children’s behavior is when the data are only consistent with one of the generalizations.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) * P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

<table>
<thead>
<tr>
<th></th>
<th>di</th>
<th>je</th>
<th>li</th>
<th>we</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>leledi</td>
<td>leleje</td>
<td>leleli</td>
<td>lelewe</td>
</tr>
<tr>
<td>wi</td>
<td>wiwidi</td>
<td>wiwije</td>
<td>wiwili</td>
<td>wiwiwe</td>
</tr>
<tr>
<td>ji</td>
<td>jijidi</td>
<td>jijije</td>
<td>jijili</td>
<td>jijiwe</td>
</tr>
<tr>
<td>de</td>
<td>dededi</td>
<td>dedeje</td>
<td>dedeli</td>
<td>dedewe</td>
</tr>
</tbody>
</table>

Control condition

Training on four word types:
leledi, wiwije, jijili, dedewe

Consistent only with the more-general hypothesis (h2).

If children fail to make the generalization in the control condition, then the results in the experimental condition will not be informative. (Perhaps the task was too hard for children.)
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Training: 2 minutes hearing artificial language words
Test: AAB pattern words using novel syllables vs.
      ABA pattern words using novel syllables

Ex: novel syllables: ko, ba
    kokoba vs.
    kobako
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Behavior: If children learn the more-general pattern (AAB), they will prefer to listen to an AAB pattern word like kokoba, over a word that does not follow the AAB pattern, like kobako.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Behavior: Children listened longer on average to test items consistent with the AAB pattern [13.51 sec], as opposed to items inconsistent with it [10.14 sec].
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

They can notice the AAB pattern and make the generalization from this artificial language data. This task isn’t too hard for infants.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

What about the experimental condition?
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Control condition
Training  leledi, wiwije, jijili, dedewe
Test      kokoba vs. kobako
Behavior  ✔

Experimental condition
Training  leledi, wiwidi, jijidi, dededi
Consistent with both a less-general hypothesis (h1) and a more-general hypothesis (h2).

\[ P(h|D) = \frac{P(D|h)\cdot P(h)}{P(D)} \]
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Control condition

Training leledi, wiwije, jijili, dedewe
Test kokoba vs. kobako

Experimental condition

Training leledi, wiwidie, jijidi, dededi
Test kokoba vs. kobako

Behavior: If children learn the more-general pattern (AAB), they will prefer to listen to an AAB pattern word like kokoba, over a word that does not follow the AAB pattern, like kobako.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Control condition
Training leledi, wiwije, jijili, dedewe
Test kokoba vs. kobako

Behavior: If children learn the less-general pattern (AAdi) or no pattern at all, they will not prefer to listen to an AAB pattern word like kokoba, over a word that does not follow the AAB pattern, like kobako.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Control condition

Training
leledi, wiwije, jijili, dedewe

Test
kokoba vs. kobako

Behavior
✔

Experimental condition

Training
leledi, wiwidi, jijidi, dededi

Test
kokoba vs. kobako

Behavior

Behavior: Children did not listen longer on average to test items consistent with the AAB pattern [10.74 sec], as opposed to items inconsistent with it [10.18 sec].
Bayesian reasoning

Gerken 2006, 2010  
Artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

Control condition

Training  leledi, wiwiye, jijili, dedewe
Test  kokoba vs. kobako
Behavior  ✔

Experimental condition

Training  leledi, wiwi, jijili, dedewe
Test  kokoba vs. kobako
Behavior  ??

They don’t learn the more-general pattern. They either learned the less-general pattern or no pattern at all.

Which one is it?
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Control condition

Training: leledi, wiwije, jijili, dedewi
Test: kokoba vs. kobako
Behavior: ✔

Experimental condition

Training: leledi, wiwidi, jijidi, dededi
Test: kokoba vs. kobako
Behavior: ??
Test: kokodi vs. kodiko
Behavior: ✔

Behavior: If they learn the less-general pattern, they’ll prefer to listen to AAdi words like kokodi.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

Control condition

Training: leledi, wiwije, jijili, dedewe
Test: kokoba vs. kobako
Behavior: ✔️

Experimental condition

Training: leledi, wiwidi, jijidi, dededi
Test: kokoba vs. kobako
Behavior: ??

Test: kokodi vs. kodiko
Behavior: ??

Behavior: If they learn no pattern at all, they’ll (again) have no preference.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

Control condition

Training leledi, wiwije, jijili, dedewe
Test kokoba vs. kobako

Behavior ✔

Children prefer to listen to novel words that follow the less-general AAdi pattern [9.33 sec] over novel words that don’t [6.25 sec].

Experimental condition

Training leledi, wiwidi, jijidi, dededi
Test kokoba vs. kobako
Behavior ??

Test kokodi vs. kodiko
Behavior ✔
Gerken 2006, 2010
artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

Control condition

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>leledi, wiwije, jijili, dedewe</td>
<td>kokoba vs. kobako</td>
<td>✔</td>
</tr>
</tbody>
</table>

Experimental condition

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>leledi, wiwidi, jijidi, dededi</td>
<td>kokoba vs. kobako</td>
<td>??</td>
</tr>
<tr>
<td>kokodi vs. kodiko</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This means that given ambiguous data, they make the less-general generalization (h1) — just like a Bayesian learner would!
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Let’s remind ourselves why this is

- **Training**
  - leledi, wiwidi, jijidi, dededi

- **Test**
  - kokodi vs. kodiko

- **Behavior** ✔
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h)*P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

<table>
<thead>
<tr>
<th></th>
<th>di</th>
<th>je</th>
<th>li</th>
<th>we</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>leledi</td>
<td>leleje</td>
<td>leleli</td>
<td>lelewe</td>
</tr>
<tr>
<td>wi</td>
<td>wiwidi</td>
<td>wiwije</td>
<td>wiwili</td>
<td>wiwiwe</td>
</tr>
<tr>
<td>ji</td>
<td>jijidi</td>
<td>jijije</td>
<td>jijili</td>
<td>jijiwe</td>
</tr>
<tr>
<td>de</td>
<td>dededi</td>
<td>dedeje</td>
<td>dedeli</td>
<td>dedewe</td>
</tr>
</tbody>
</table>
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

\[ \propto P(D|h) \times P(h) \]

likelihoods
\[ P(D | h_1) = 1/4 \times 1/4 \times 1/4 \times 1/4 = 1/256 \]

These are the only 4 data that can be generated, and so the probability of generating each one is 1/4. Let’s focus on the types in the data intake, so we just have these four.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h)*P(h)}{P(D)} \]

\(\propto P(D|h) * P(h)\)

Likelihoods
\[ P(D \mid h_1) = 1/256 \]
\[ P(D \mid h_2) = 1/16*1/16*1/16*1/16 \]
\[ = 1/65536 \]

These are 16 data that can be generated, and so the probability of generating each one is 1/16.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]
\[ \propto P(D|h) \times P(h) \]

likelihoods
\[ P(D \mid h_1) = 1/256 \]
\[ P(D \mid h_2) = 1/65536 \]

priors
Let’s assume the hypotheses are equally complex a priori, so they have uniform prior probability.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$
\[\propto P(D|h) \cdot P(h)\]

likelihoods
P(D | h1) = 1/256
P(D | h2) = 1/65536

priors
P(h1) = 1/2
P(h2) = 1/2
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h | D) = \frac{P(D | h) \cdot P(h)}{P(D)} \]
\[ \propto P(D | h) \cdot P(h) \]

**likelihoods**
- \( P(D | h1) = 1/256 \)
- \( P(D | h2) = 1/65536 \)

**priors**
- \( P(h1) = 1/2 \)
- \( P(h2) = 1/2 \)

**posteriors**
- \( P(h1 | D) \propto 1/256 \cdot 1/2 \)
- \( P(h2 | D) \propto 1/65536 \cdot 1/2 \)
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]
\[ \propto P(D|h) \cdot P(h) \]

likelyhoods
- \( P(D | h_1) = 1/256 \)
- \( P(D | h_2) = 1/65536 \)

priors
- \( P(h_1) = 1/2 \)
- \( P(h_2) = 1/2 \)

posteriors
- \( P(h_1 | D) \propto 1/256 \cdot 1/2 \)
- \( P(h_2 | D) \propto 1/65536 \cdot 1/2 \)

\( h_1 \) is 256 times \( (1/256 \text{ vs. } 1/65536) \) as probable as \( h_2 \)

Therefore, prefer \( h_1 \).
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[
P(h|D) = \frac{P(D|h) \times P(h)}{P(D)}
\]
\[\propto P(D|h) \times P(h)\]

**likelihoods**

- P(D | h1) = 1/256
- P(D | h2) = 1/65536

**priors**

- P(h1) = 1/2
- P(h2) = 1/2

**posteriors**

- P(h1 | D) \propto \frac{1}{256} \times \frac{1}{2}
- P(h2 | D) \propto \frac{1}{65536} \times \frac{1}{2}

Note how it’s the likelihood doing all the work.

**Therefore, prefer h1.**
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

\[ \propto P(D|h) \cdot P(h) \]

Another important point:
Bayesian learners are sensitive to counterexamples.
Bayesian reasoning

**Gerken 2006, 2010**

Artificial language study

\[ P(h|D) = \frac{P(D|h)P(h)}{P(D)} \]

\[ \propto P(D|h)P(h) \]

Sensitive to **counterexamples**

If even one word in the intake wasn’t compatible with the less-general **AAdi** pattern, a Bayesian learner would notice that and shift beliefs.
Bayesian reasoning

**Gerken 2006, 2010**
artificial language study

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

sensitive to **counterexamples**

If even one word in the intake **wasn’t compatible** with the less-general **AA\text{di}** pattern, a Bayesian learner would notice that and shift beliefs.

Why? This has to do with the **likelihood**.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h)P(h)}{P(D)} \]

sensitive to counterexamples

likelihood
\[ P(D | h1) = 1/4 \times 1/4 \times 1/4 \times 1/4 \times 0 = 0 \]

These are the only 4 data that can be generated, and so the probability of generating each one is 1/4 except the last one, which can’t be generated.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]
\[ \propto P(D|h) \times P(h) \]
sensitive to **counterexamples**

**likelihood**
\[ P(D \mid h_1) = 0 \]
\[ P(D \mid h_2) = \frac{1}{16} \times \frac{1}{16} \times \frac{1}{16} \times \frac{1}{16} \times \frac{1}{16} = \frac{1}{1048576} \]

In contrast, even though the other data points have a smaller probability of being generated by \( h_2 \), the last one *can* be generated, so the likelihood isn’t 0.
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h | D) = \frac{P(D|h) \times P(h)}{P(D)} \]

sensitive to counterexamples

likelihood
\[ P(D | h1) = 0 \]
\[ P(D | h2) = 1/1048576 \]

This means only \( h2 \) will have a non-zero posterior, and so the Bayesian learner prefers \( h2 \).
Bayesian reasoning

Gerken 2006, 2010
artificial language study

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

sensitive to counterexamples

Do 9-month-olds reason this way too?
Bayesian reasoning

Gerken 2006, 2010 sensitive to **counterexamples**
artificial language study

Task type: Head Turn Preference Procedure
with 9-month-olds

**Training** leledi, wiwidi, jijidi, dededi + 3 AAB examples (like lelewe)

2 minutes a few seconds at the end

\[ P(h|D) = \frac{P(D|h)*P(h)}{P(D)} \]
Bayesian reasoning

$P(h|D) = \frac{P(D|h)*P(h)}{P(D)}$

Gerken 2006, 2010 sensitive to counterexamples
artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

Training  leledi, wiwidi, jijidi, dededi + 3 AAB examples (like lelewe)
          2 minutes

Test     kokoba vs. kobako
Bayesian reasoning

Gerken 2006, 2010 sensitive to counterexamples artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

**Training** leledi, wiwidi, jijidi, dededi + 3 AAB examples (like lelewe) 2 minutes a few seconds at the end

**Test** kokoba vs. kobako

**Behavior** ✔

Behavior: If they learn the more-general pattern from these three counterexamples, they’ll prefer to listen to AAB words like kokoba.
Bayesian reasoning

Gerken 2006, 2010 sensitive to counterexamples artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

Training leledi, wiwidi, jijidi, dededi + 3 AAB examples (like lelewe) 2 minutes

Test kokoba vs. kobako

Behavior ✔

Children prefer to listen to novel words that follow the more-general AAB pattern [~11 sec] over novel words that don’t [~8 sec]
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \]

**Gerken 2006, 2010**

sensitive to **counterexamples**

artificial language study

Task type: Head Turn Preference Procedure with 9-month-olds

*Training*  leledi, wiwidi, jijidi, dededi + 3 AAB examples (like lelewe)

2 minutes

*Test*  kokoba vs. kobako

*Behavior*  ✔

This is noticeably different than their behavior when they only hear AAdi examples in their intake.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h)\cdot P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Takeaway: At 9 months, infants show probabilistic reasoning abilities similar to a Bayesian learner.
Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) * P(h)}{P(D)} \]

Gerken 2006, 2010
artificial language study

Takeaway: At 9 months, infants show probabilistic reasoning abilities similar to a Bayesian learner.

When given ambiguous data compatible with two hypotheses, a less-general and more-general one, they choose the less-general one (which gives a higher likelihood to the data).
Bayesian reasoning

Takeaway: At 9 months, infants show probabilistic reasoning abilities similar to a Bayesian learner.

- ambiguous data = less-general hypothesis

When given even a very few counterexamples that are only compatible with the more-general hypothesis, they shift their beliefs accordingly.

$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$
Today’s Plan:
Bayesian inference & linguistic parameters

I. Bayesian reasoning

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]

II. Parameters & overhypotheses

III. Structure dependence
Today’s Plan:
Bayesian inference & linguistic parameters

II. Parameters & overhypotheses
Parameters & overhypotheses

Remember:
We can think of grammars as collections of parameter values.
Parameters & overhypotheses

A parameter (and its specific value) determines what we predict will be observed in the world in a variety of situations.
A parameter determines what we predict will be observed. Example: Head-directionality

Linguistic parameters correspond to the properties that vary across human languages.
A parameter determines what we predict will be observed.

The fact that parameters connect to multiple structural properties is a very good thing for acquisition. This is because a child can learn about that parameter’s value by observing many different kinds of examples in the language.
A parameter determines what we predict will be observed.

**good for acquisition**

Let’s assume a number of properties are all connected to parameter $P$, which can take one of two values: a or b.
A parameter determines what we predict will be observed.

**good for acquisition**

Let’s assume a number of **properties** are all connected to parameter $P$, which can take one of two values: $a$ or $b$. 

- property 1
- property 2
- property 3

$P = a$ or $b$?
A parameter determines what we predict will be observed.

Head-directionality

good for acquisition
How do we learn whether property 3 shows behavior a or b? One way is to observe instances of property 3 in the intake.
A parameter determines what we predict will be observed.

Head-directionality

**good for acquisition**

But what if property 3 occurs very rarely? We might never see any examples of property 3.

\[ P = a \text{ or } b? \]

property 1

property 2

property 3

???
A parameter determines what we predict will be observed.

**Parameters & overhypotheses**

**good for acquisition**
Fortunately, because property 3 is connected to P, we can learn the value for property 3 by learning the value of P.
Parameters & overhypotheses

A parameter determines what we predict will be observed.

Grammar

Head-directionality

good for acquisition
Also fortunately, P is connected to properties 1 and 2.

P = a or b?

property 1

property 2

property 3
A parameter determines what we predict will be observed.

**good for acquisition**
This means we can learn the value of $P$ from property 1 or property 2.
A parameter determines what we predict will be observed.

Head-directionality

good for acquisition
Suppose we see an example of property 1 with value a.
A parameter determines what we predict will be observed.

**good for acquisition**
This means P also should have value \textit{a}.
A parameter determines what we predict will be observed.

**good for acquisition**

So, we can make predictions for all the other properties connected to P, even if we’ve never seen examples of them. This is great!
A parameter determines what we predict will be observed.

**good for acquisition**
This highlights another benefit - we don’t have to learn the behavior of each structure individually.
A parameter determines what we predict will be observed.

**good for acquisition**
Instead, we can observe some properties (like property 1) and infer the right behavior for the remaining properties (like property 2 and property 3).
A parameter determines what we predict will be observed.

**good for acquisition**
That is, instead of having to make 3 decisions (one for properties 1, 2, and 3), we actually only need to make one decision - is \( P \) a or b?
Parameters & overhypotheses

A parameter determines what we predict will be observed.  

Head-directionality

good for acquisition
The intake is used to make this one decision, which generates useful predictions for other properties of the language.
Overhypotheses in hierarchical Bayesian learning are generalizations made at a more abstract level, which cover many different data types.

In this way, they’re similar in spirit to linguistic parameters.
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

Suppose you’re observing the contents of marble bags.
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

The first bag you look at has 20 black marbles.
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

The second bag you look at has 20 white marbles.

20
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

linguistic parameter

20 20
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

The third and fourth bags you look at have **20 black marbles**.
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

You get a fifth bag and pull out a single marble. It’s white.
What do you predict about the color distribution of the rest of the marbles in the bag?
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

Probably that they’re all white!
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

What if you then get another bag and pull out a single purple marble from it? What would you predict?
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

Probably that all the rest of the marbles in the bag are purple, too!
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

Why does this happen?
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

It seems like you’re learning something about the color distribution *in general* (not just for a particular bag): all marbles in a bag have the same color.
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

This allows you to make predictions when you’ve only seen a single marble of whatever color from a bag.
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

overhypothesis
all the same color

all black  all white  all black  all black
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

overhypothesis
all the same color

all black
all white
all black
all black
all something

20 20 20 20
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

overhypothesis
all the same color

all black  all white  all black  all black  all purple

20  20  20  20  1
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

overhypothesis
all the same color

all black all white all black all black

all purple

Seem familiar?
Parameters & overhypotheses

Overhypotheses

Non-linguistic example

overhypothesis
all the same color

all black
all white
all black
all black
all purple

Seem familiar?
Bayesian learning models are able to learn overhypotheses, provided they know **what the parameters are and the range of values those parameters can take.**

(ex: Kemp, Perfors, & Tenenbaum 2007, Perfors, Tenebaum, & Wonnacott 2010).
Parameters & overhypotheses

Bayesian learning models are able to learn overhypotheses, provided they know what the parameters are and the range of values those parameters can take.

What about real learners (children)?
When provided with partial evidence about a few objects in a few categories, can infants form a more abstract generalization (an overhypothesis) that then applies to a new category?

Dewar & Xu 2010
9-month-olds
Training trials:
Observe four different objects pulled out by experimenter who had her eyes closed - the objects are different colors but always have the same shape.
Dewar & Xu 2010
9-month-olds

Training: different colors but same shape

Experimental condition
If infants create an overhypothesis that all objects in a box have the same shape...
Dewar & Xu 2010

9-month-olds

Training: different colors but same shape

Experimental condition
If infants create an overhypothesis that all objects in a box have the same shape...

they should expect the experimenter to pull out all the same shape from a new box.

This shouldn’t be surprising, and so infants shouldn’t look as long at it.
Parameters & overhypotheses

Dewar & Xu 2010
9-month-olds

Training: different colors but same shape

**Experimental condition**
If infants create an overhypothesis that all objects in a box have the same shape...

...they shouldn’t expect the experimenter to pull out different shapes from a new box, even if one is a shape they’ve seen before.

This should be surprising, and so infants should look longer at it.
Dewar & Xu 2010  
9-month-olds

Training: different colors but same shape

Control condition

If infants create an overhypothesis that all objects in a box have the same shape...

they should expect the experimenter to pull out different shapes from different boxes.

This shouldn’t be surprising, and so infants shouldn’t look as long at it.

Note how this outcome looks identical to the experimental condition outcome.
Parameters & overhypotheses

Dewar & Xu 2010
9-month-olds

Training: different colors but same shape

Control condition
If infants create an overhypothesis that all objects in a box have the same shape...
they should expect the experimenter to pull out different shapes from different boxes.

This shouldn’t be surprising, and so infants shouldn’t look as long at it.

The only difference is how the outcome was generated (from the same box or from different boxes — which is what the overhypothesis is about).
If infants create an overhypothesis that all objects in a box have the same shape.

This is what we expect.
Dewar & Xu 2010
9-month-olds

Training:

different colors but same shape

If infants create an overhypothesis that all objects in a box have the same shape

Experimental condition

~11.32 sec 😞

~14.28 sec 😞

And this is exactly what happened!

Control condition

~10.3-11.0 sec 😞
Parameters & overhypotheses  

**Dewar & Xu 2010**

9-month-olds

**Training:**

If infants create an **overhypothesis** that all objects in a box have the **same shape**

**Experimental condition**

~11.32 sec 😞

~14.28 sec 😞

9-month-olds appear able to form overhypotheses from very limited data sets.

**Control condition**

~10.3-11.0 sec 😞
If infants create an overhypothesis that all objects in a box have the same shape, Dewar & Xu (2010) reported that 9-month-olds took about 11.32 seconds to complete the training task. The control condition, which involved different colors but the same shape, took about 14.28 seconds. Hopefully, this means they can also use linguistic parameters to learn, since parameters are similar to overhypotheses about language!
Today’s Plan:
Bayesian inference & linguistic parameters

I. Bayesian reasoning

\[ P(h|D) = \frac{P(D|h)P(h)}{P(D)} \]

II. Parameters & overhypotheses

III. Structure dependence
Today’s Plan: Bayesian inference & linguistic parameters

III. Structure dependence
Idea: Rules for word order depend on linguistic structure.
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Statement
Jareth can alter time.

How do we turn this into a question whose answer is either yes or no?
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Yes/No question
Can Jareth alter time?

What changed?
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Statement
Jareth can alter time.

Yes/No question
Can Jareth alter time?

Where the auxiliary can appears.
Where the noun/subject Jareth appears.
Structure dependence
Rules for word order **depend on linguistic structure**

An example: Yes/No question formation in English

**Statement**

Jareth *can* alter time.

**Yes/No question**

*Can* Jareth *alter* time?

Where the auxiliary *can* appears.

Where the noun/subject *Jareth* appears.

The child’s job: Figure out the rule for turning statements into yes/no questions.
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Rule: Something about one or both of these?
Where the auxiliary *can* appears.
Where the noun/subject *Jareth* appears.

Rule? Swap the order of the first two words
Rule? Swap the order of the subject and the auxiliary
Rule? Move the first noun to the second position
Rule? Move the auxiliary to the first position

And there are others...

Let’s look at some additional data.
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

This one doesn’t capture the pattern.

Rule? Swap the order of the first two words
Rule? Swap the order of the subject and the auxiliary
Rule? Move the first noun to the second position
Rule? Move the auxiliary to the first position
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

Which auxiliary and what’s “swapping” mean if they’re not next to each other?

Rule? Swap the order of the subject and the auxiliary
Rule? Move the first noun to the second position
Rule? Move the auxiliary to the first position
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

This doesn’t handle “would” being in the first position.

Rule? Move the first noun to the second position
Rule? Move the auxiliary to the first position
Structure dependence

Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

Which auxiliary?

Rule? Move the auxiliary to the first position
Structure dependence
Rules for word order depend on linguistic structure
An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

This would capture the first question’s pattern too.

Rule? Move the last auxiliary to the first position

Let’s look at some additional data.
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.
       Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.
       Would anyone who can wish away their brother be tempted to do it?

Someone who can solve the labyrinth can show someone else who can’t how.
       Can someone who can solve the labyrinth show someone else who can’t how?

This doesn’t capture the pattern.

Rule? Move the last auxiliary to the first position

Now what?
Structure dependence

Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

Someone who can solve the labyrinth can show someone else who can’t how.

Can someone who can solve the labyrinth show someone else who can’t how?

This doesn’t capture the pattern.

Rule? Move the last auxiliary to the first position

Let’s try incorporating structure.
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

Someone who can solve the labyrinth can show someone else who can’t how.

Can someone who can solve the labyrinth show someone else who can’t how?

Rule? Move the main clause auxiliary to the first position
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.
Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.
Would anyone who can wish away their brother be tempted to do it?

Main subject
Someone who can solve the labyrinth can show someone else who can’t how.
Can someone who can solve the labyrinth show someone else who can’t how?

✔ Rule? Move the main clause auxiliary to the first position
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.
Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.
Would anyone who can wish away their brother be tempted to do it?

Main subject
Someone who can solve the labyrinth can show someone else who can’t how.

Main objects
Can someone who can solve the labyrinth show someone else who can’t how?

Rule? Move the main clause auxiliary to the first position
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

Main subject

Someone who can solve the labyrinth can show someone else who can’t how.

Can someone who can solve the labyrinth show someone else who can’t how?

Rule? Move the main clause auxiliary to the first position
Structure dependence

Rules for word order **depend on linguistic structure**

An example: Yes/No question formation in English

**Jareth can alter time.**

**Can Jareth alter time?**

**Anyone who can wish away their brother would be tempted to do it.**

**Would anyone who can wish away their brother be tempted to do it?**

**Someone who can solve the labyrinth can show someone else who can’t how.**

**Can someone who can solve the labyrinth show someone else who can’t how?**

✔ Rule? **Move the main clause auxiliary to the first position**

This also works for the other examples.
Structure dependence
Rules for word order depend on linguistic structure

An example: Yes/No question formation in English

Jareth can alter time.

Can Jareth alter time?

Anyone who can wish away their brother would be tempted to do it.

Would anyone who can wish away their brother be tempted to do it?

Someone who can solve the labyrinth can show someone else who can’t how.

Can someone who can solve the labyrinth show someone else who can’t how?

✔ Rule? Move the main clause auxiliary to the first position

Because this rule refers to clause structure, it’s structure-dependent.
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English

✔ Rule? Move the main clause auxiliary to the first position

When do children figure this out?
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English

✔ Rule? Move the main clause auxiliary to the first position

Crain & Nakayama 1987
Elicited productions from three- to five-year-olds
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English

✔ Rule? Move the main clause auxiliary to the first position

Crain & Nakayama 1987
Elicited productions from three- to five-year-olds

“Ask Jabba if...

“...the boy who can see Mickey Mouse is happy.”
“...the boy who is happy can see Mickey Mouse.”
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English

✔ Rule? Move the main clause auxiliary to the first position

Crain & Nakayama 1987
Elicited productions from three- to five-year-olds

Common errors that occurred:

(Restarts)
- simplifying the subject so main clause auxiliary is more accessible
“Is the boy who can see Mickey Mouse, is he happy?”
“Can the boy who is happy, can he see Mickey Mouse?”

“Ask Jabba if...

“...the boy who can see Mickey Mouse is happy.”
“...the boy who is happy can see Mickey Mouse.”
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English

☑ Rule? Move the main clause auxiliary to the first position

Crain & Nakayama 1987
Elicited productions from three- to five-year-olds

Common errors that occurred:
(Restarts) - simplifying the subject so main clause auxiliary is more accessible
(Initial is prefix) - giving up (sort of a generic question marking)
“Is the boy who can see Mickey Mouse is happy?”
“Is the boy who is happy can see Mickey Mouse?”

“Ask Jabba if...

“...the boy who can see Mickey Mouse is happy.”
“...the boy who is happy can see Mickey Mouse.”
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English
✔ Rule? Move the main clause auxiliary to the first position

Crain & Nakayama 1987
Elicited productions from three- to five-year-olds

Common errors that occurred:
(Restarts) - simplifying the subject so main clause auxiliary is more accessible
(Initial is prefix) - giving up (sort of a generic question marking)

Errors that didn't occur (Structure-independent auxiliary movement)
“Can the boy who __ see Mickey Mouse is happy?”
“Is the boy who __ happy can see Mickey Mouse?”

“Ask Jabba if...

“...the boy who can see Mickey Mouse is happy.”
“...the boy who is happy can see Mickey Mouse.”
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English

✅ Rule? Move the main clause auxiliary to the first position

Crain & Nakayama 1987
Elicited productions from three- to five-year-olds

Common errors that occurred:
(Reverts) - simplifying the subject so main clause auxiliary is more accessible
(Initial is prefix) - giving up (sort of a generic question marking)

Errors that didn't occur (Structure-independent auxiliary movement)

How we can interpret this: As young as three years old, children have some very specific constraints on the kind of hypotheses they’ll consider for complex yes/no questions.
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

How could they learn this?
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

A potential input issue

Most of the yes/no question data children encounter (particularly before the age of 3) consists of simple yes/no questions compatible with many different rules.

Rule? Swap the order of the first two words
Rule? Swap the order of the subject and the auxiliary
Rule? Move the first noun to the second position
Rule? Move the auxiliary to the first position
Rule? Move the main clause auxiliary to the first position
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

A potential input issue

Most of the yes/no question data children encounter (particularly before the age of 3) consists of simple yes/no questions compatible with many different rules.

But structure-dependence is a very general property about language...
Structure dependence
Rules for word order depend on linguistic structure

Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

A potential input issue
Most of the yes/no question data children encounter (particularly before the age of 3) consists of simple yes/no questions compatible with many different rules.

It could be an overhypothesis about language.

Jareth can alter time.
Can Jareth alter time?
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

A potential input issue

Most of the yes/no question data children encounter (particularly before the age of 3) consists of simple yes/no questions compatible with many different rules.

And this overhypothesis would connect to many other structures besides yes/no questions.

Jareth can alter time.

Can Jareth alter time?

The girl in the Labyrinth thought something.

What did the girl in the Labyrinth think?
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

A potential input issue

Most of the yes/no question data children encounter (particularly before the age of 3) consists of simple yes/no questions compatible with many different rules.

And this overhypothesis would connect to many other structures besides yes/no questions.
Structure dependence
Rules for word order depend on linguistic structure

Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

A potential input issue - may not be as bad
Children could encounter a lot of data that might favor structured representations over unstructured ones (e.g., linear structures)

overhypothesis

Jareth can alter time.
Can Jareth alter time?

The girl in the Labyrinth thought something.
What did the girl in the Labyrinth think?
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

A potential input issue - may not be as bad

Children could encounter a lot of data that might favor structured representations over unstructured ones (e.g., linear structures)

overhypothesis

prefer structured representations

Jareth can alter time.

Can Jareth alter time?

The girl in the Labyrinth thought something.

What did the girl in the Labyrinth think?
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

computational-level modeled learner
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Learned from realistic samples of child-directed English speech

“\textit{I love kitties.}”

Lidz & Gagliardi 2015
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Learned from realistic samples of child-directed English speech abstracted into syntactic category sequences
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Hypotheses

There are different types of grammars available (e.g., structure-dependent vs. linear)
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Hypotheses

grammar type

There are specific grammars of each type (e.g., different structure-dependent grammars)
Structure dependence

Rules for word order depend on linguistic structure.

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Hypotheses

grammar type

specific grammar

Each grammar connects to specific structures in the observable data.
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Grammar type
Specific grammar
Structures in observable data

Use Bayesian inference to infer the best grammar type & specific grammar, given the child-directed speech data.

\[ P(h|D) = \frac{P(D|h)*P(h)}{P(D)} \]
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Grammar type

Specific grammar

Structures in observable data

Note: The priors for different grammars aren’t equal. Structure-dependent grammars are more complex than other grammar types being considered, and so have lower prior probability.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

This means structure-dependent grammars are actually disfavored a priori!
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Note: The priors for different grammars aren’t equal. Structure-dependent grammars are more complex than other grammar types being considered, and so have lower prior probability.

\[
P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}
\]

This means they really have to do a better job accounting for the data to be preferred!
Structure dependence
Rules for word order depend on linguistic structure
Yes/No question formation in English
By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Grammar type
structure-dependent

Specific grammar
structures in observable data

And this is exactly what happens!

\[ P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \]
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

grammatical type
structure-dependent

specific grammar

structures in observable data

Even for the earliest child-directed speech samples (directed at children two years old), the structure-dependent grammar types are preferred.
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

**grammar type**
structure-dependent

**specific grammar**

**structures in observable data**

Why? Because many different data types favor structure-dependent representations over other simpler representations.
Structure dependence

Rules for word order depend on linguistic structure

Yes/No question formation in English

By three years old, children have some very specific structure-dependent constraints on hypotheses about word order.
Structure dependence
Rules for word order depend on linguistic structure

By three years old, children have some very specific constraints on hypotheses about word order.

Perfors, Tenenbaum, & Regier 2011

Yes/No question formation in English
And so these structure-dependent representations make hypothesizing structure-dependent rules much more probable.
Thank you!

Lisa S. Pearl
Associate Professor
Department of Linguistics
Department of Cognitive Sciences
SSPB 2219, SBSG 2314
University of California, Irvine
lpearl@uci.edu

\[ P(h|D) = \frac{P(D|h)P(h)}{P(D)} \]